A Technique To Help Choose Between Alternative Functional Forms Of The Regression Equation

by

Thomas P. Frazier
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria, Virginia 22311

Presented at the
19th Annual Department of Defense
Cost Analysis Symposium
Xerox Training Center
Leesburg, Virginia
September 17 - 20 1985

The Cost Analysis Symposium was sponsored by: OSD (PA&E) Resource Analysis, Cost Analysis Division, Pentagon, Washington, DC 20301

A TECHNIQUE TO HELP CHOOSE BETWEEN ALTERNATIVE FUNCTIONAL FORMS OF THE REGRESSION EQUATION

Introduction

The cost analyst is often faced with the problem of choosing among several different functional forms of a regression equation. For example, a problem frequently encountered in empirical cost research is the choice between a linear and a log-linear equation as shown below:

$$Y = a_1 + b_1X + c_1Z + U_1 \qquad [1]$$

$$LnY = a_2 + b_2\,LnX + c_2\,LnZ + U_2 \qquad [2]$$

Ideally, theory tells the researcher which to choose. However, in the absence of any firm theoretical indication, the cost researcher must resort to some ad hoc procedure in choosing between the functional forms. If the appropriate functional form is not used the estimate may be biased and/or inefficient.

A procedure often used is to run both functional forms and select the one with the largest coefficient of determination (R-square). Unfortunately, the R-square statistic of equations [1] and [2] cannot be meaningfully compared because the dependent variables in the two equations are different (i.e. Y in equation [1] and LnY after transformation in equation [2]). Rao and Miller state, "The specification of the model, the error terms, and the computation of R-square for these two equations are entirely different and provide no common ground for comparison of the relative performance of these equations on the basis of computation of R-square."[1]

[1] P. Rao and R.L. Miller, Applied Econometrics (Belmont, Calif: Wadsworth, 1971), pp. 18.

Transformation of the Data

Fortunately, a technique developed by Box and Cox allows the researcher to bring these different equations onto a common footing.[1] This standardizing procedure involves putting the dependent variable data into comparable dimensionless units by transforming the Y data by

$$Y^* = kY \qquad [3]$$

where

$$k = \exp\ ((-\Sigma \text{Ln } Y)/N) \qquad [4]$$

is the inverse of the geometric mean of Y. If we multiply each observation by k then [1] and [2] may be expressed as

$$Y^* = a_1^* + b^*_1 X + c^*_1 Z + U^*_1 \qquad [5]$$

$$\text{Ln}Y^* = a_2^* + b^*_2 \text{Ln}X + c^*_2 \text{Ln}Z + U_2^* \qquad [6]$$

Since the residual sums of squares in equations [5] and [6] are directly comparable, the cost researcher can now choose the functional form possessing the minimum residual sum of squares as the appropriate functional form.

Box and Cox also developed a nonparametric test to examine whether the difference between the residual sums of squares of these two functional forms is significant. The test statistic d is defined as

$$d = \frac{N}{2} \left| \text{Ln} \frac{\Sigma e^{*2}_1}{\Sigma e^{*2}_2} \right| \qquad [7]$$

[1] G.E.P. Box and D.R. Cox, "An Analysis of Transformation, "Journal of the Royal Statistical Society, Series B, 1964, 211-143. This subject is also discussed by P. Rao and R.L. Miller in Applied Econometrics, pp. 107-111.

where

$$\Sigma e^{*2}_1 \quad \text{and} \quad \Sigma e^{*2}_2$$

are the residual sums of squares in equations [5] and [6] respectively. The d statistic is assumed to follow a Chi-squared distribution with one degree of freedom. The null hypothesis is specified as the two functional forms being empirically equivalent. If the d statistic exceeds the critical value, the cost researcher may reject the null hypothesis of equivalency.

Examples

Two examples using actual data have been calculated to illustrate the Box-Cox procedure. The first example involves the estimation of a production function using cross-sectional data with thirteen observations. The second example involves the estimation of the cost of Navy Stock Account repair parts procured by the ship for use in maintenance of the ship and installed equipment. A cross-sectional sample of 47 surface ships was used in the regression.

Example 1

This example is taken from a study undertaken in connection with a comprehensive evaluation of new investments in the nation's Air Traffic Control (ATC) System [1]. It involved a quantitative analysis of the relationship between ATC system outputs and inputs. Outputs are defined in terms of operations handled while inputs consist of labor and capital. The data presented below consist of a cross-section of observations on individual radar approach towers in a single year.

[1]H. Eskew, T. Frazier, and M. Smith, "An Econometric Analysis of Enroute and Terminal Air Traffic Control," ASC-R-110, June 1976.

The original linear and log-linear functions are

$$Y = -72280 + 4114.04\,L + 24.96\,K \qquad [8]$$
$$\qquad\qquad (477.8) \qquad (7.7)$$

$$\Sigma e_1^2 = 11780559872$$

$$LnY = 6.14 + 0.78\,LnL + 0.38\,LnK \qquad [9]$$
$$\qquad\qquad (0.11) \qquad (0.16)$$

$$\Sigma e_2^2 = .47$$

where $\Sigma\, e_1^2$ and $\Sigma\, e_2^2$ are the residual sums of squares in these two equations, respectively, and Y represents number of operations, L the number of controllers, and K the amount of capital. The standard errors are presented in the parenthesis.

Since the dependent variables in the above equations are not the same, the sums of residual squares are not directly comparable. Using the transformation discussed previously, we can make them comparable. The inverse of the geometric mean of Y is .000005; thus k = .000005. Transforming the dependent variables in [8] and [9] yields $Y^* = .000005\ Y$.

The new (transformed) residual sums of squares of equations [8] and [9] now become

$$Y = -0.361 + 0.020\,L + 0.0012\,K \qquad [8a]$$
$$\qquad\qquad (0.002) \quad (0.00003)$$

$$\Sigma e_1^{*2} = 0.2945$$

$$LnY = -6.06 + 0.78\,LnL + 0.38\,LnK \qquad [9a]$$
$$\qquad\qquad (0.11) \qquad (0.16)$$

$$\Sigma e_2^{*2} = .4700$$

Since $\Sigma e_1^{*2} < \Sigma e_2^{*2}$, the linear functional form [8a] appears empirically more appropriate than the log-linear form for the production function.

Computing the d statistic to determine whether these two functions are empirically equivalent by substituting the appropriate data into [6] yields

$$d = \frac{13}{2} \left| \text{Ln} \frac{0.2945}{0.4700} \right| = 3.038$$

The critical value for the 95 percent level of confidence at one degree of freedom is 3.84. The computed statistic, 3.038 is less than the critical value. Hence we accept the null hypothesis that these functions are empirically equivalent with 95% confidence. Thus, although the linear functional form yields a smaller residual sum of squares than does the log-linear form, in this case it does not matter (from a statistical point of view) which functional form is selected.

Example 2

In the preceding example, we saw that although the linear form produced a smaller (corrected) residual sum of squares than the log-linear form, the difference between the two functional forms was not significant. In this example, we will demonstrate a situation in which the two functional forms are signficantly different in explanatory power.

This example is taken from a study of the operating and support costs of Naval ships.[1] The initial linear and log-linear functions are

$$Y = \underset{}{-156.10} + \underset{(0.04)}{0.07}\ SHU + \underset{(0.006)}{0.007}\ SHP \qquad [10]$$

$$\Sigma e_1^2 = 3{,}845{,}794$$

$$LnY = \underset{}{-5.09} + \underset{(0.17)}{0.55}\ LnSHU + \underset{(0.06)}{0.61}\ LnSHP \qquad [11]$$

$$\Sigma e_2^2 = 10.87$$

[1] T.P. Frazier, "Naval Ships Operating and Support Cost Handbook," ASC R-122, Dec. 1979.

where Y is the annual cost of repair parts for a ship in thousands of fiscal 1978 dollars, SHU represents the total steaming hours per year underway, and SHP is the total shaft horsepower of the ship.

Using the Box-Cox transformation, we can make [10] and [11] comparable so as to select a preferred form. The inverse of the geometric mean of Y is .00358 and substituting the transformed values of the dependent variable into equations [10] and [11] produces the following pair of regression equations with corresponding estimated standard errors and residual sums of squares:

$$Y = \underset{}{0.56} + \underset{(0.0011)}{0.0025}\ SHU + \underset{(0.00002)}{0.0002}\ SHP \qquad [10a]$$

$$\Sigma e_1^{*2} = 49.54$$

$$LnY = \underset{}{-10.72} + \underset{(0.17)}{0.55}\ Ln\ SHU + \underset{(0.06)}{0.61}\ Ln\ SHP \qquad [11a]$$

$$\Sigma e_2^{*2} = 10.87$$

In this case the log-linear form produces a smaller residual sum of squares than the linear model. Computing the d statistic to determine whether these two functions are empirically equivalent yields

$$d = \frac{47}{2}\left| Ln\ \frac{49.54}{10.87} \right| = 35.64$$

which is significantly larger than the critical value of 3.84. Thus we reject the null hypothesis that these functions are empirically equivalent with 95% confidence and would select the log-linear form.

Sample Size and the d Statistic [1]

As the two examples suggest, the sample size plays an important role in determining whether to accept or reject the null hypothesis of equivalence. Indeed, as Figure 1 illustrates, as the sample size starts to get larger, almost any difference in the two sums of squares becomes significant. The top half of Figure 1 traces the ratio of the transformed linear residual sum of squares to the long-linear value. For example, if the sample size (N) is 10 and the ratio is greater than 2 (1.83 to be exact), the linear functional form is rejected in favor of the log-linear. If the sample size is 50, the ratio need only be 1.12 before the linear functional form is rejected.

The bottom half of Figure 1 is the mirror image of the top half except that it measures the reciprocal of

$$\frac{\Sigma e^{*2}_1}{\Sigma e^{*2}_2}$$

Expressing it this way gives symmetry above and below the 1.0 line. In this case with N=10, the log-linear residual sum of squares is more than 1.83 the linear value, the log-linear functional form is rejected in favor of the linear.

The point of this discussion is to alert the analyst to the usefulness of the transformation and hpothesis test when he is working with relatively large samples.

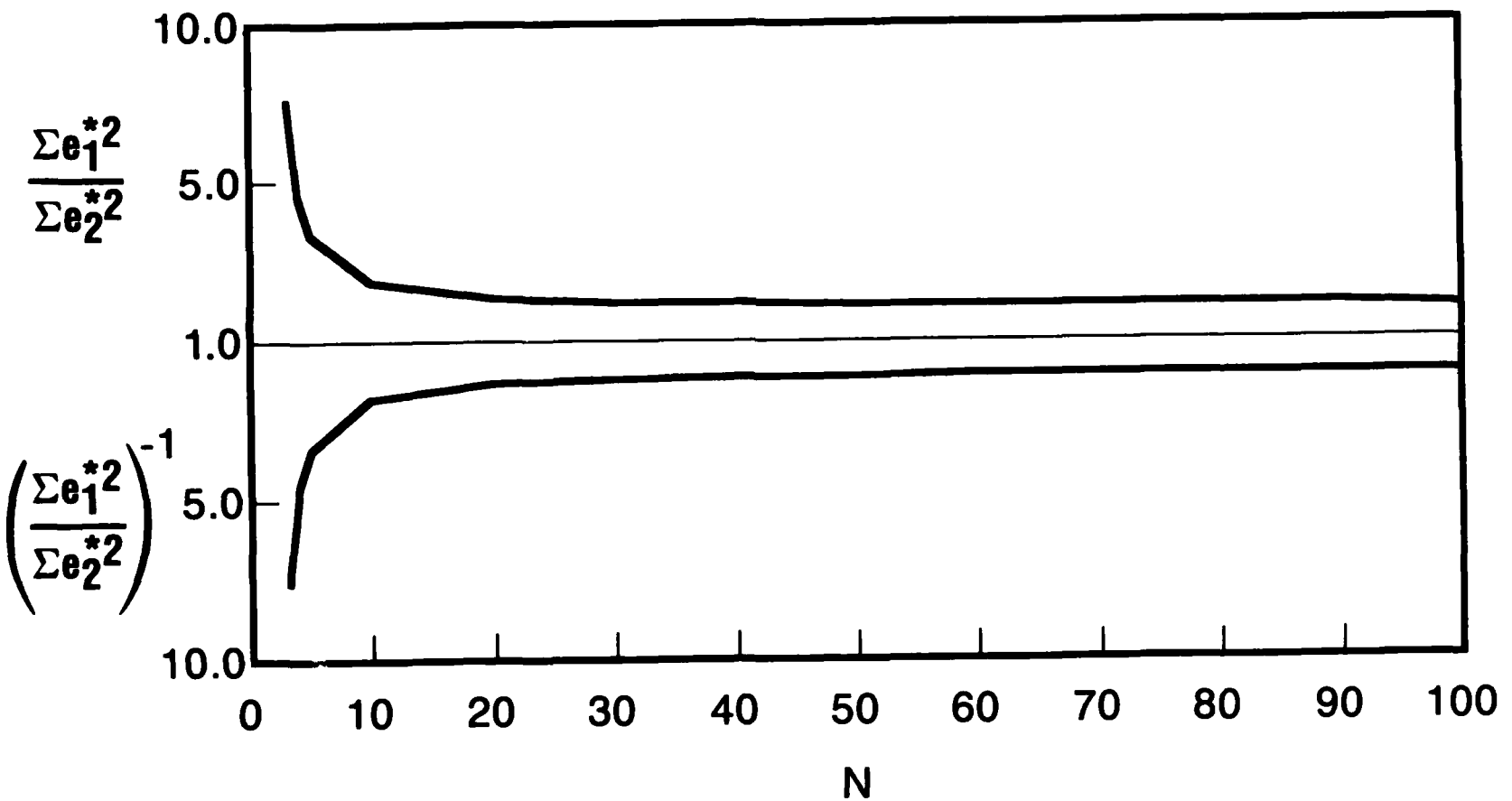
Summary

This paper has presented a technique that may be of some help to cost analysts in selecting between two functional forms of a regression equation. The technique, developed by Box and Cox, involves the transformation of the dependent variable. The resultant transformed

[1] My thanks to Mr. Henry Eskew of the Center for Naval Analyses for bringing this topic to my attention and also for suggesting Figure 1.

MINIMUM VALUES OF $\frac{\Sigma e_1^{*2}}{\Sigma e_2^{*2}}$ AND $\left(\frac{\Sigma e_1^{*2}}{\Sigma e_2^{*2}}\right)^{-1}$ FOR REJECTING H_o

AS A FUNCTION OF SAMPLE SIZE



equations are directly comparable for the purpose of selecting the most appropriate form. A nonparametric test to examine whether the difference between the two functional forms is significant is also detailed. Two examples using actual data are provided in order to illustrate the usefulness of the technique. The effect of the size of the sample on the nonparametric test is also discussed.